

Titre de la thèse : **Intégration de données dans un système question-réponse sur le Web**

Véronique Moriceau

Soutenue le 15 février 2007 à l'IRIT,
Devant le jury composé de :

- Guy LAPALME Professeur à l'université de Montréal – RALI (*rapporteur*)
- Pierre ZWEIGENBAUM Directeur de recherche CNRS – LIMSI (*rapporteur*)
- Régine ANDRÉ-OBRECHT Professeur à l'université Paul Sabatier – IRIT (*présidente*)
- Christian RETORÉ Professeur à l'université Bordeaux I - LaBRI
- Patrick SÉGUÉLA Synapse Développement
- Patrick SAINT-DIZIER Directeur de recherche CNRS – IRIT (*directeur de thèse*)

Résumé :

Aujourd'hui, le Web met à la disposition du grand public un très grand nombre de données et les systèmes de recherche d'informations développés ces dernières années sont des outils pratiques pour qui souhaite trouver une réponse à une question sur tout type de domaine. L'interrogation de ces moteurs de recherche se fait sur la base de mots-clés et ceux-ci proposent à l'utilisateur un ensemble de liens vers des pages Web et/ou des extraits de ces pages traitant du thème de la requête. En revanche, là où les dictionnaires, encyclopédies et bases de données fournissent une réponse synthétique et cohérente, une analyse rapide des réponses proposées par les moteurs montrent que celles-ci sont très souvent non pertinentes (car elles ne répondent pas à la question posée), incohérentes, etc. C'est donc à l'utilisateur qu'il revient de trier et de rechercher au sein de ces pages la réponse à sa question. Cette démarche entraîne une perte de temps considérable, sans pour autant avoir l'assurance de trouver une réponse correcte. Le problème qui se pose alors est le **problème de la pertinence et de la cohérence des réponses** proposées à l'utilisateur. Les systèmes question-réponse proposent une alternative à ces problèmes.

Pour cette thèse, nous nous plaçons dans le cadre des systèmes question-réponse coopératifs sur le Web. Nos principaux objectifs sont de modéliser, concevoir et évaluer un système capable, à partir d'une question en langue naturelle, de rechercher les réponses pertinentes sur le Web et de générer en langue naturelle une réponse synthétique, même quand le moteur de recherche sélectionne plusieurs réponses potentielles. Pour cela, il faut intégrer entre autres les attentes des utilisateurs et des techniques de traitement de l'information. Travaillant en domaine ouvert, nous nous sommes plus particulièrement intéressés à deux types de questions : les questions attendant des réponses temporelles (dates) et des réponses numériques.

Notre système a donc pour but :

- l'intégration d'informations provenant des différentes réponses potentielles à une question en utilisant une base de connaissances et des connaissances directement extraites des pages Web. Ce composant permet notamment de détecter les incohérences de données et de prendre en compte les attentes de l'utilisateur pour produire une réponse appropriée,
- la production en langue naturelle de réponses synthétiques et pertinentes vis-à-vis de l'utilisateur. La coopérativité se situe à plusieurs niveaux : il faut produire des réponses courtes, intelligibles et qui expriment le savoir-faire coopératif mis en oeuvre pour résoudre les incohérences de données,

Nous proposons également des méthodes d'évaluation adéquates pour évaluer le système :

- d'un point de vue technique en évaluant les performances de chacun des composants du système,
- d'un point de vue cognitif en confrontant les utilisateurs aux réponses produites.

Mots-clés : système question-réponse, Web, intégration de données, génération de langue